

Sashuv Kafle

kafle.sashuv@gmail.com ❖ (+977) 9840031140 ❖ Kathmandu, Nepal ❖ GitHub ❖ Website

Research interests: trustworthy and efficient foundation models, uncertainty quantification, vision–language and multimodal models, knowledge distillation, diffusion models, and reinforcement learning.

EDUCATION

Kathmandu University, School of Science

2021 – 2026

B.Sc. in Computational Mathematics

Kathmandu, Nepal

- **Relevant coursework:** Probability & Statistics, Stochastic Processes, Numerical Methods, Linear Algebra, Optimization, Data Structures & Algorithms.

RESEARCH PROJECTS

Fixation of Dominated Strategies in Group-Normalized Policy Gradient

Independent research

2026 – present

- Modelled RLVR/GRPO training as a two-strategy selection process and showed that the drift and the sampling noise of the update carry the same state-dependent factor, so their ratio stays constant along the whole run.
- Derived a closed-form fixation probability (classical gambler’s ruin) controlled by a single dimensionless parameter combining group size, learning rate and selection strength.
- Validated it in a vectorised Python/NumPy simulator over 45 configurations (mean absolute error 0.006, all collapsing onto one curve): a strictly worse strategy took over permanently in ~31% of runs at small group size, invisible to benchmark accuracy.

Severity Aware Conformal Risk Control For Medical Claim Filtering

Independent research

*Under Review at **ML4H***

- First-author work extending distribution-free risk guarantees to tier-specific risk budgets, holding clinically dangerous claims to a stricter bound than benign ones.
- Ran a four-way score ablation (self-confidence, sequence log-probability, NLI entailment, random) showing the guarantee holds for any scoring function.

Limiting Sycophancy in Large Language Models via Conformal Risk Control

Independent research

*Under Review at **MLHC***

- Introduced Conformal Claim Refinement (CCR): decomposes an LLM response into atomic claims, scores claim validity with a judge model, applies claim-level conformal risk control to filter high-risk statements, then rewrites the response from those that survive.
- Evaluated sycophancy across multiple model families and parameter scales on MedQuAD and HealthSearchQA, covering regressive and progressive variants; CCR cut sycophancy by up to 7.4% on average with no retraining.

EXPERIENCE

Nepal Applied Mathematics and Informatics Institute for Research (NAAMII)

Sept 2025 – Present

Research Intern

Kathmandu, Nepal

- Investigate the factuality and reliability of large language models, developing post-hoc methods that reduce hallucination and sycophancy with distribution-free guarantees via conformal prediction and conformal risk control.
- Lead first-author work on uncertainty quantification for black-box LLMs in safety-critical healthcare settings, where model internals are unavailable and guarantees must come from sampled outputs alone.

Khalti Digital Wallet

Apr 2026 – Jul 2026

AI Engineer

Kathmandu, Nepal

- Built and deployed retrieval-augmented generation (RAG) pipelines for financial question answering and customer support at one of Nepal’s leading payment platforms, focusing on grounding and safe model outputs in a regulated domain.

SELECTED PROJECTS

- **Visualizing RLVR Entropy Collapse** — reproduces entropy collapse in RL with verifiable rewards and the covariance mechanism behind it. [GitHub](#)
- **RL Game Agent with Proximal Policy Optimization** — Tic-Tac-Toe agent trained from scratch with PPO, implemented end to end. [GitHub](#)
- **Partition-Assisted K-Nearest Neighbors** — KNN variant partitioning the search space to cut query cost while preserving accuracy. [GitHub](#)

TEACHING & SERVICE

- **Teaching Assistant**, Annual Nepal AI School (ANAIS) — assisted lectures and workshops on machine learning; prepared course material and evaluated assignments.
- **Technical Head**, Computational Mathematics Club, Kathmandu University — led technical initiatives and built computational tools for club projects and student research.

TECHNICAL SKILLS

Languages & tools: Python, JavaScript, SQL, LaTeX, Git, Linux, React Native

ML & numerics: PyTorch, TensorFlow, Hugging Face, NumPy, SciPy, scikit-learn

Methods: Conformal prediction & risk control, uncertainty quantification, RL (PPO, GRPO), stochastic simulation, RAG